Elbow Method In K Means Clustering

SHUBHAM PATEL

BATCH - 03

SERIAL NO. 104

Contents

Clustering

An unsupervised learning method.

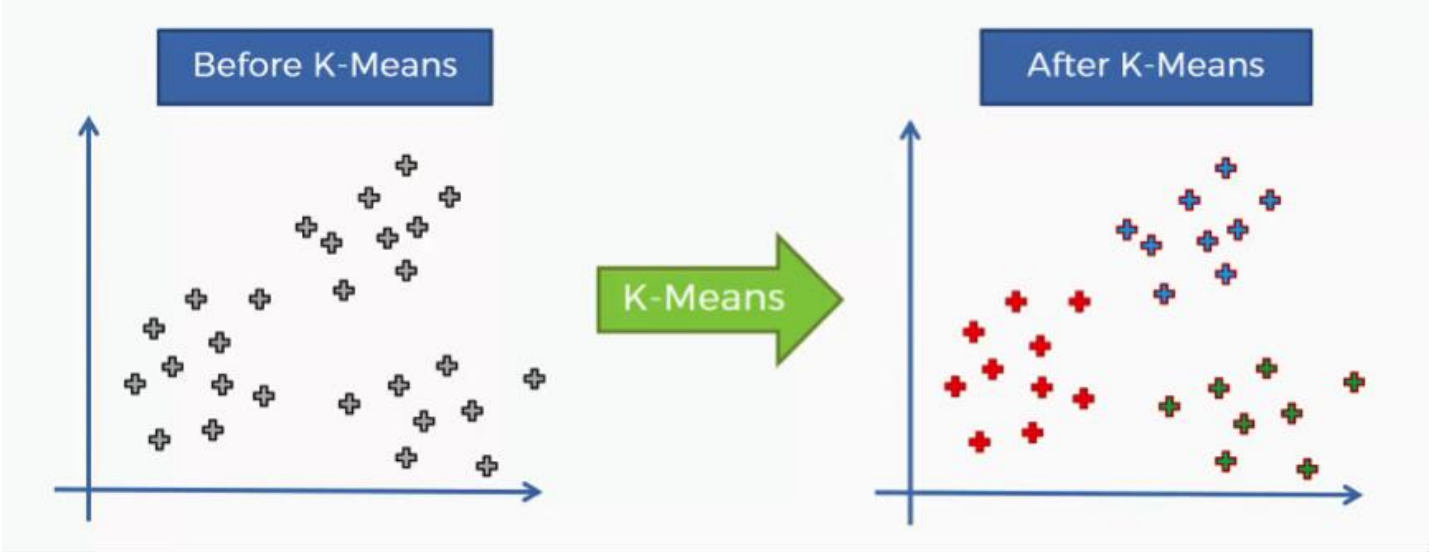
No predefined labels to cluster objects.

Types of Clustering:

- K Means clustering
- K Medoids clustering
- DBSCAN (Density bases special clustering of applications and noise)
- OPTICS (Ordering Points to identify Clustering structure)
- Hierarchical Clustering

K Means Clustering

K-means is a clustering algorithm which is centroid-based or we can say is based on the center point of each cluster, along with distance-based or we can say the distance between the cluster center explained above and points in a cluster , where we calculate the distances to assign a point to a cluster. In K-Means, each cluster is represented by a centroid or the cluster mean.



Few important parameters of sklearn.cluster.KMeans()

from sklearn.cluster import KMeans

n_clusters: int, default=8
The number of clusters to form as well as the number of centroids to generate.

init: {‘k-means++’, ‘random’}, callable or array-like of shape (n_clusters, n_features), default=’k-means++’
Method for initialization

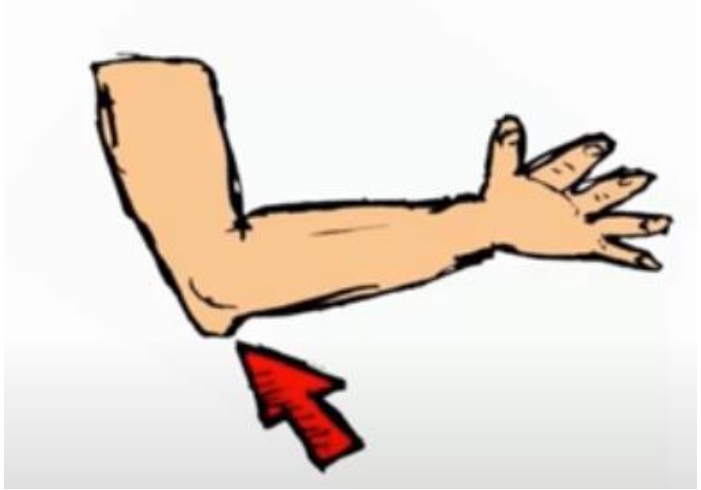
n_init: int, default=10
Number of time the k-means algorithm will be run with different centroid seeds. The final results will be the best output of n_init consecutive runs.

max_iter: int, default=300
Maximum number of iterations of the k-means algorithm for a single run.

random_state: int, RandomState instance or None, default=None
Determines random number generation for centroid initialization.

Elbow Method

The elbow method helps to choose the optimum value of ‘k’ (number of clusters) by fitting the model with a range of values of ‘k’. We calculate the sum of squared distance(SSD) of each object from its cluster center, and from that, we find the number until which sharp changes in SSD are occurring. Through this, are finding the number of clusters through which even after increasing the number of clusters, does not make a significant difference or in other words, we are finding clusters for which increasing cluster centers does not bring significant difference in SSD. This point seems like an elbow and is the required number of k we want, and thus the method is named as elbow method.



```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
import numpy as np
from sklearn.cluster import KMeans
```

```
[2] data=pd.read_csv('california_cities.csv')
    data.head(10)
```

	city	latd	longd	population_total
0	Adelanto	34.576111	-117.432778	31765
1	AgouraHills	34.153333	-118.761667	20330
2	Alameda	37.756111	-122.274444	75467
3	Albany	37.886944	-122.297778	18969
4	Alhambra	34.081944	-118.135000	83089
5	AlisoViejo	33.575000	-117.725556	47823
6	Alturas	41.487222	-120.542500	2827
7	AmadorCity	38.419444	-120.824167	185
8	AmericanCanyon	38.168056	-122.252500	19454
9	Anaheim	33.836111	-117.889722	336000

Implementing Elbow Method for K Means clustering

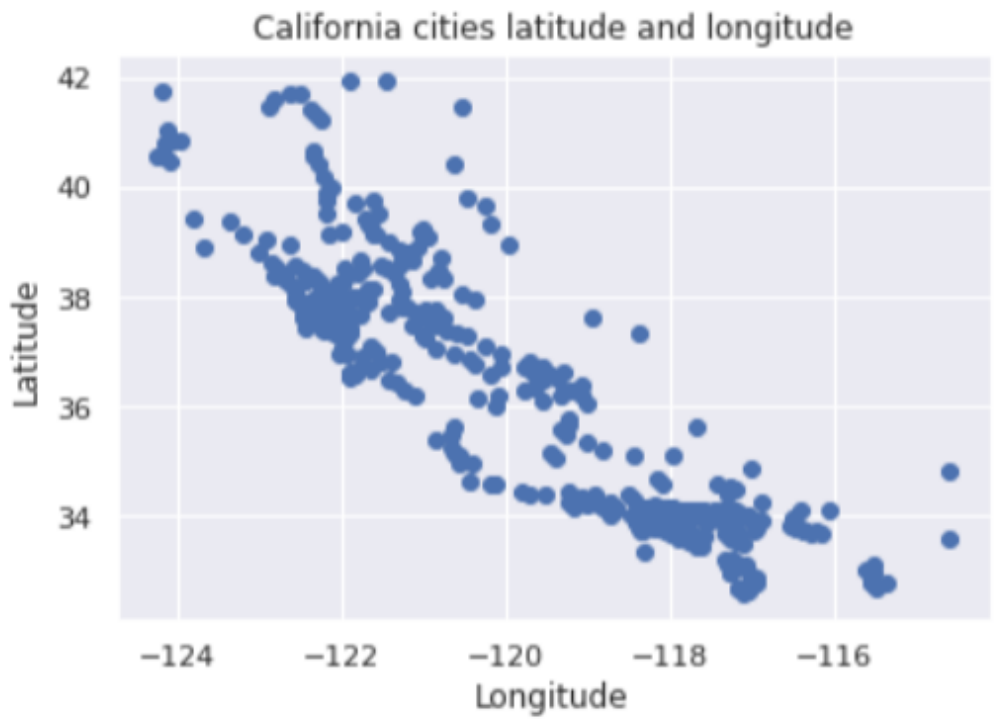
Importing required libraries and reading the dataset

Unclustered data

```
x=data.iloc[:,1:3]
x.head()
```

	latd	longd
0	34.576111	-117.432778
1	34.153333	-118.761667
2	37.756111	-122.274444
3	37.886944	-122.297778
4	34.081944	-118.135000

```
#Plotting scatter plot for raw data of longitude and latitude
plt.scatter(data['longd'],data['latd'])
plt.ylabel('Latitude')
plt.xlabel('Longitude')
plt.title('California cities latitude and longitude')
plt.show()
```



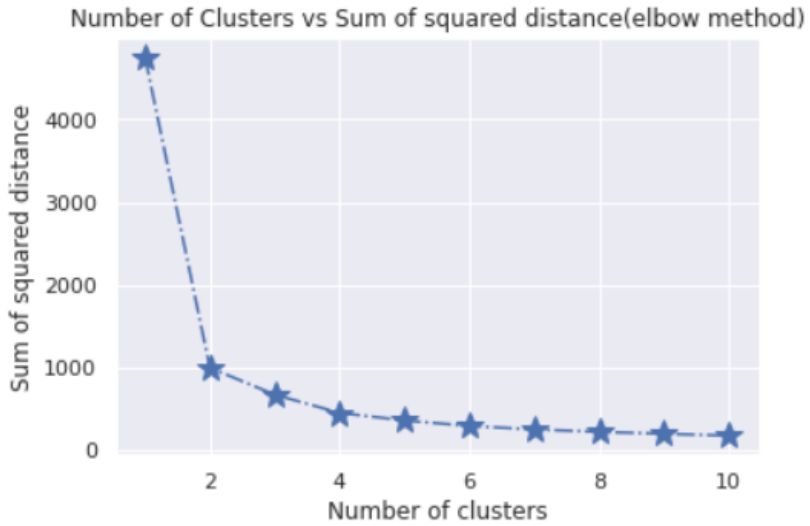
Elbow method Implementation

```
#Using elbow method to find optimal number of clusters
ssd=[]
for i in range(1,11):
    kmeans=KMeans(n_clusters=i,init='k-means++',max_iter=300,n_init=10)
    kmeans.fit(x)
    ssd_iter=kmeans.inertia_
    ssd.append(ssd_iter)
```

```
print(ssd)
```

```
[4758.6989431177235, 994.7028222089918, 667.2440520115674, 447.9038159329313, 355.91001445846985, 2
```

```
number_clusters=range(1,11)
plt.plot(number_clusters,ssd,marker='*',linestyle='-.',markersize=15)
plt.title('Number of Clusters vs Sum of squared distance(elbow method)')
plt.xlabel('Number of clusters')
plt.ylabel('Sum of squared distance')
plt.show()
```



Advantage and Disadvantage of Elbow method

- Advantage: Through this method, we are able to find the optimum number of clusters, which is difficult task in K Means clustering algorithm and this particular task of finding number of clusters is considered to be a disadvantage of K Means algorithm, which is overcomed by elbow method.
- Disadvantage: The presence of multidimensional data and/or outliers affect this elbow method same as K Means clustering algorithm. Because of this, uneven clustering takes place.

Thankyou